Implementasi *Web Scraping* dan *Text Mining* untuk Akuisisi dan Kategorisasi Informasi Laman *Web* Tentang Hidroponik

A Priyanto[1], M R Ma'arif[2*]

[1] Program Studi Teknik Informatika, Universitas Jenderal Achmad Yani Yogyakarta
[2] Program Studi Sistem Informasi, Universitas Jenderal Achmad Yani Yogyakarta

Email: agungpriyanto@gmail.com[1], muhammad.rifqi@gmail.com[2*]

Masuk: 11 Juni 2018, direvisi: 13 Agustus 2018, diterima: 26 Agustus 2018

Abstrak. Dengan banyaknya sumber informasi yang ada, akan memunculkan dua kemungkinan, di satu sisi akan memberikan manfaat, namun di sisi lain akan menimbulkan fenomena *information overload*. *Information overload* adalah banyaknya jumlah informasi yang diterima oleh manusia sehingga menimbulkan kesulitan dalam penerimaan dan pengolahan. Fenomena *information overload* salah satunya terjadi pada informasi mengenai tata cara bercocok tanam dengan metode atau teknik hidroponik yang sekarang sedang marak digemari masyarakat luas. Dengan banyaknya laman web yang menyajikan informasi mengenai hidroponik, masyarakat harus menyediakan lebih banyak waktu untuk memilah dan mengakses sebanyak mungkin laman *web* guna mendapatkan informasi yang lengkap dan akurat. Penelitian ini bertujuan untuk mengimplementasikan teknik *web scraping* yang dikombinasikan dengan *text mining* untuk secara otomatis mengakuisisi informasi dari laman-laman *web* yang memuat informasi mengenai hidroponik dan mengkategorisasikannya berdasarkan topik yang lebih spesifik dari artikel hidroponik yang terdapat dalam laman *web* tersebut. Dari eksperimen yang sudah dilakukan, *web scraping* dan *text mining* berhasil diimplementasikan untuk mengakuisisi artikel-artikel terkait hidroponik dari internet dan mengelompokkannya ke dalam beberapa kategori berdasarkan topik artikel secara otomatis.

Kata Kunci: *Web Scraping*, *Text Mining*, Akuisisi Informasi, Kategorisasi Informasi, Hidroponik

Abstract. A huge amount of information throughout internet gives a lot of benefits. On the other hand, it can also cause a phenomenon called information overload. Information overload can be defined as a difficulty faced by the human to process information due to the amount of information they accept. This phenomenon can occur on a lot of aspects in life. One of them is hydroponic procedures. Hydroponic is one of many farming methods that nowadays is widely adopted by citizens especially those who have no background as a farmer due to its portability and convenience. On the internet, citizens can find a lot of web pages about hydroponic, hence they have to provide much time to filter all of them in order to obtain complete and accurate information. Thus, this experiment is aimed to implement two techniques named web scraping combined with text mining. Web scraping technique was used to acquire the information from web pages on the internet, whereas text mining was used to categorize the information based on its topic. As a result, a significant number of web pages about hydroponic were successfully acquired and automatically categorized into several topics.

Keywords: Web Scraping, Text Mining, Information Acquisition, Information Categorization, Hydroponic

1. Pendahuluan

Saat ini dengan penetrasi internet yang semakin masif di masyarakat, berbagai informasi bisa didapatkan dengan sangat mudah dalam waktu yang relatif singkat. Banyaknya sumber informasi yang ada di satu sisi akan memberikan manfaat, karena antara satu sumber dengan sumber lainnya dapat saling melengkapi informasi yang disajikan. Namun, di sisi lain informasi dengan jumlah yang sangat banyak dan beragam akan menyebabkan timbulnya *information overload. Information overload* adalah banyaknya informasi yang diterima oleh manusia sehingga sulit untuk mengolahnya. Karena adanya *information overload*, manusia dituntut untuk dapat mengombinasikan berbagai informasi yang didapatkan dari berbagai sumber sehingga menjadi satu kesatuan informasi yang utuh, akurat dan bermanfaat. Salah satu komponen penting dalam manajemen pengetahuan adalah akuisisi informasi. Komponen ini berperan penting dalam memilih dan memilah informasi yang relevan untuk dimasukkan ke dalam struktur pengetahuan yang akan dikembangkan. Karakteristik informasi yang tersedia di internet memiliki perbedaan yang cukup signifikan dengan karakteristik informasi yang tersedia di dalam suatu organisasi baik dari sisi bentuk, variasi maupun validitas informasi. Perbedaan ini tentunya memerlukan pendekatan yang berbeda dalam proses akuisisinya dibandingkan dengan proses akuisisi informasi di internal organisasi. Dalam penelitian ini akan dikembangkan sebuah metodologi akuisisi informasi dari laman *web* yang tersedia di internet. Metodologi yang dibangun akan memanfaatkan teknologi *web scraping* dan *text mining*.

Penelitian ini bertujuan untuk mengeksplorasi laman-laman yang menyajikan informasi mengenai tata cara bercocok tanam dengan teknik mengumpulkan semua informasi yang ada di laman tersebut secara otomatis dengan menggunakan teknologi *web scraping*. Informasi yang sudah terkumpul kemudian akan dikelompokkan dengan menggunakan teknologi *text mining* sehingga mempermudah pengguna dalam hal ini pehobi hidroponik untuk menemukan artikel/tutorial berdasarkan tema/topik tertentu. Studi kasus hidroponik ini diambil atas dua pertimbangan. Pertimbangan yang pertama adalah banyaknya laman *web* mengenai pengelolaan dan tata cara bertanam dengan metode hidroponik, namun masih sangat sedikit laman web yang menyajikan informasi secara lengkap dan sistematis. Pertimbangan yang kedua adalah, saat ini hidroponik merupakan alternatif metode tanam yang digemari masyarakat karena kepraktisannya. Sehingga, dengan pengembangan manajemen pengetahuan untuk pengelolaan tanaman hidroponik, masyarakat dapat memperoleh informasi yang lebih lengkap dan terstruktur.

Penggunaan teknik *web scraping* untuk mengambilan data dari laman web secara otomatis sudah dikenal luas. Dalam penelitian yang dilakukan oleh Wijaya [1] teknik *web scraping* digunakan untuk membuat portal pencarian produk *smartphone* dengan menggabungkan informasi dari laman-laman *web e-commerce*. Contoh kasus lain dari pemanfaatan *web scraping* dilakukan oleh Ma'arif [2]. Dalam penelitian tersebut *web scraping* digunakan untuk mengumpulkan informasi mengenai objek-objek wisata di Daerah Istimewa Yogyakarta kemudian menggabungkannya ke dalam satu portal informasi terpadu. Penggunaan *web scraping* dengan teknik lebih lanjut dilakukan oleh Kadam [3] serta Dastidar [4]. Kadam memanfaatkan *web scraping* untuk mengambil dan membandingkan harga *spare part* komputer. Sementara Dastidar menggunakan *web scraping* untuk mengumpulkan informasi mengenai portofolio seseorang di internet untuk keperluan *profiling* dan *surveillence*.

Dari beberapa penelitian yang penulis uraikan di atas, semuanya hanya sebatas memanfaatkan *web scraping* untuk mengambil dan menggabungkan data dari laman-laman *web* yang ada kemudian menyajikannya secara mentah kepada pengguna. Dalam kajian ini, teknik *web scraping* akan digabungkan dengan teknik *text mining*, sehingga nantinya informasi yang dikumpulkan akan diolah terlebih dahulu untuk disajikan kedalam bentuk atau format yang lebih singkat dan terstruktur kepada pengguna tanpa mengurangi muatan informasi yang disajikan.

2. Kerangka Teoritis

2.1. Web Scraping

Web scraping adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa laman *web* yang dibangun dengan bahasa *markup* seperti HTML atau XHTML yang bertujuan untuk mengambil informasi dari halaman tersebut baik secara keseluruhan atau sebagian untuk digunakan bagi kepentingan lain [5]. Secara umum, ada empat tahapan dalam penggunaan *web scraping* untuk mengambil data secara otomatis dari sebuah laman *web* sebagai berikut [6]:

1. Mempelajari dokumen HTML dari *website* yang akan diambil informasinya untuk *tag* HTML yang mengapit informasi yang akan diambil.
2. Menelusuri mekanisme navigasi pada *website* yang akan diambil informasinya untuk ditirukan pada aplikasi *web scraper* yang akan dibuat.
3. Berdasarkan informasi yang didapat pada langkah 1 dan 2 di atas, aplikasi *web scraper* dibuat untuk mengotomatisasi pengambilan informasi dari *website* yang ditentukan.
4. Informasi yang didapat dari langkah 3 disimpan dalam format data tertentu.

Dalam penelitian ini *web scraping* digunakan untuk mengambil data dari sebuah laman *web* kemudian melakukan transformasi dari bentuk yang tidak terstruktur, umumnya dalam format HTML menjadi suatu format data terstruktur yang dapat disimpan ke dalam *database* untuk keperluan repositori maupun analisis lebih lanjut.

2.2. Text Mining

Text mining adalah sebuah teknik/pendekatan algoritmik berbasis komputer untuk mendapatkan suatu pengetahuan baru yang tersembunyi dari sekumpulan teks. *Text mining* merupakan bagian dari keilmuan *information retrieval* (temu balik informasi) yang bekerja pada data bertipe teks yang cenderung tidak terstruktur [7]. Pada dasarnya mekanisme kerja algoritma-algoritma *text mining* memiliki kemiripan dengan algoritma-algoritma *data mining* secara umum. Perbedaan pokok dari *text mining* dan *data mining* adalah pada tipe data yang menjadi objek kerjanya [8]. Jika *data mining* bekerja pada data terstruktur yang maka *text mining* bekerja pada data yang tidak terstruktur. Kombinasi antara *text mining* dan *data mining* dapat digunakan untuk menyelesaikan masalah-masalah klasifikasi, klastering, maupun prediksi pada informasi yang bersifat tekstual.

Seperti halnya *data mining*, *text mining* merupakan pendekatan algoritmik yang secara sistematis memproses data teks melalui beberapa tahapan. Secara umum, tahapan besar dalam *text mining* terdiri dari tiga bagian utama yakni *text preprocessing*, *feature selection*, dan *text analytic* [9]. Penjelasan lebih lanjut dari tahap-tahap tersebut adalah sebagai berikut :

1. *Text Preprocessing.*
   Tahapan ini adalah tahapan yang berfungsi untuk membersihkan teks sebelum diolah lebih lanjut. Data teks mentah yang tidak terstruktur memiliki cukup banyak *noise* seperti tanda baca, angka, imbuhan, karakter-karakter khusus, *slang word* dan lain sebagainya. Dalam tahapan ini, data teks tersebut dibersihkan sehingga tersisa bentuk dasarnya saja untuk keperluan analisis teks lebih lanjut.
2. *Feature Selection.*
   Tahapan ini berperan dalam menentukan *term*/kata kunci yang menjadi ciri dari suatu dokumen yang membedakan dokumen tersebut dengan dokumen yang lain dalam satu korpus. Dalam *text mining*, feature selection merupakan tahapan yang paling penting yang memiliki peran yang sangat signifikan dalam akurasi *text analytic*. Empat pendekatan yang paling umum digunakan dalam *feature selection* adalah *Document Frequency* (DF), *Term Frequency* (TF), *Inverse Document Frequency* (IDF) dan *Term Frequency/Inverse Document Frequency* (TF/IDF).
   a. *Document Frequency* (DF).
      Prinsip kerja dari DF adalah membuang *term-term* yang umum terdapat di dokumen-dokumen yang ada pada suatu korpus dokumen teks. Sehingga *term* yang tersisa

dalam suatu dokumen adalah *term-term* yang memiliki tingkat *overlapping* yang rendah dengan *term-term* yang terdapat di dokumen lain dalam suatu korpus.

b. *Term Frequency* (TF).
Berbeda dengan DF, pendekatan TF tidak mengindahkan *term* yang terkandung dalam dokumen lain. Metode TF hanya secara sederhana menghitung kemunculan *term* dalam suatu dokumen. *Term-term* yang memiliki frekuensi kemunculan tinggi akan menjadi ciri dari suatu dokumen dimana *term* tersebut berada.

c. *Inverse Document Frequency* (IDF).
Pendekatan IDF mirip dengan TF, yakni menghitung frekuensi kemunculan suatu term. Namun, jika TF menghitung kemunculan suatu *term* hanya di satu dokumen teks, maka IDF menghitung kemunculan suatu term di keseluruhan korpus dokumen.

d. *Term Frequency/Inverse Document Frequency* (TF/IDF).
TF/IDF adalah gabungan dari pendekatan TF dan IDF dengan mengambil rasio antara nilai TF dan nilai IDF.

3. *Text Analytic*.
Tahapan terakhir dari proses *text mining* adalah *text analytic*. Dalam tahapan ini data teks yang sudah dibersihkan dan diidentifikasikan berdasarkan *term*/kata kunci yang menjadi ciri dokumen teks tersebut diolah dengan menggunakan berbagai macam algoritma untuk berbagai kebutuhan analisis. Dua jenis *text analytic* yang paling sering dilakukan adalah *topic modelling* dan *sentiment analysis*. *Topic modelling* adalah sebuah pendekatan untuk mengelompokkan teks/dokumen teks kedalam beberapa kategori secara otomatis berdasarkan tingkat kesamaan *term*/kata kunci. Sedangkan *sentiment analysis* adalah sebuah pendekatan untuk mengestimasi/mengklasifikasikan teks ke dalam berbagai macam *sentiment* (positif, negative, netral, sarkas, dan lain sebagainya).

3. Metodologi

Penelitian ini adalah rancang bangun sebuah program komputer untuk mengakuisisi data dari laman web yang bias diakses secara bebas dan mengelompokkan data tersebut sesuai dengan topik bahasannya. Rancang bangun program ini dikembangkan dengan metode *waterfall*. Metode ini dipilih karena kebutuhan-kebutuhan terkait pengembangan sistem sudah bisa diidentifikasi di tahapan awal pengembangan. Metode *waterfall* terdiri dari empat tahapan yaitu [10] :

1. Tahapan identifikasi dan analisis kebutuhan.
Dalam tahapan ini peneliti akan mengidentifikasi struktur serta karakteristik konten dari laman-laman *web* yang ada dan pengolahan data dengan menggunakan *web scraping* dan *text mining*.
2. Tahapan desain sistem.
Dalam tahapan ini peneliti akan menentukan desain program *web scraping* dan *text mining* yang tepat untuk mengambil dan mengolah data secara otomatis dari setiap laman *web*.
3. Tahapan implementasi sistem.
Dalam tahapan ini peneliti akan mengimplementasikan desain program yang dibuat pada tahap kedua ke dalam bahasa pemrograman.
4. Tahapan pengujian sistem.
Dalam tahapan ini peneliti akan melakukan pengujian terhadap hasil penelitian. Pengujian dilakukan dalam dua tahapan. Tahap yang pertama adalah pengujian performa sistem yang dibangun dalam melakukan *scraping* terhadap laman-laman *web* yang lain, dan tahapan yang kedua adalah pengujian tingkat akurasi *text mining* dalam mengolah data yang diambil.

Pendekatan khusus yang digunakan dalam penelitian ini adalah teknik *web scraping* untuk akuisisi data dan *text mining* untuk kategorisasi data artikel hidroponik yang didapatkan dari laman *web* yang bisa diakses secara bebas. Untuk keperluan akuisisi data dengan *web scraping*, maka sebagai langkah awal adalah menentukan laman-laman *web* yang akan diakusisi. Dari hasil penelusuran dengan menggunakan mesin pencari Google, diambil 10 situs yang akan dijadikan objek dalam

eksperimen ini. Daftar situs yang digunakan dalam penelitian ini diperlihatkan oleh tabel 1. Secara lebih spesifik kriteria yang digunakan untuk menentukan situs yang digunakan dalam eksperimen adalah sebagai berikut:

1. Merupakan situs yang secara khusus membahas mengenai hidroponik, atau
2. Situs yang memiliki halaman khusus yang membahas mengenai hidroponik.

Setelah penentuan laman web yang akan diakusisi, langkah selanjutnya adalah melakukan ekstraksi informasi dari laman sumber menggunakan teknik *web scraping*. Dalam penelitian ini, digunakan pustaka *Scrapy* yang ditulis dalam bahasa pemrograman *Python* untuk mempermudah proses *scraping*.

Tabel 1. Daftar situs tentang hidroponik yang menjadi target *webscraping*.

No.	URL
1	http://belajarberkebun.com/category/hidroponik
2	http://hidroponikshop.com/blog
3	http://agroteknologi.web.id/page/%s/?s=Hidroponik
4	https://goodplant.co.id/blog/
5	http://www.mediahidroponik.com/
6	http://hidropedia.com/blog/
7	http://www.kebunhidro.com/
8	http://tanamtanaman.com/category/hidroponik/
9	http://hidroponikpedia.com
10	http://mitalom.com/category/hidroponik/

Informasi/artikel hidroponik yang berhasil diakuisisi selanjutnya dikelompokkan dengan menggunakan teknik *text mining*. Dalam penelitian ini, proses *text mining* terbagi kedalam dua tahapan, yakni *text pre-processing* dan *topic modelling*. Tahapan *text mining* yang pertama adalah *text pre-processing*. Dalam melakukan *text mining*, teks dokumen yang digunakan harus dipersiapkan terlebih dahulu, setelah itu baru dapat digunakan untuk proses utama. Proses mempersiapkan teks dokumen atau *dataset* mentah disebut juga dengan proses *text preprocessing. Text preprocessing* berfungsi untuk membersihkan data teks yang didapatkan dari internet. Dalam eksperimen ini tahapan *preprocessing* yang dilakukan mengadopsi tahapan *preprocessing* dalam eksperimen yang dilakukan oleh Hidayatullah dan Ma'arif [11] dalam mengolah kalimat berbahasa Indonesia, dengan beberapa modifikasi. Lebih lanjut, ada tiga tahapan *text preprocessing* yang dilakukan sebagai berikut :

1. Tokenisasi, yaitu pemotongan kalimat berdasarkan kata yang menyusunnya.
2. *Stopword removal*, yaitu tahapan membuang kata-kata yang tidak berpengaruh terhadap proses klasifikasi seperti kata depan, kata sambung, dan lain sebagainya.
3. *Case folding*, yaitu tahapan untuk menyeragamkan bentuk huruf menjadi huruf besar atau huruf kecil.

Tahapan kedua dari proses *text mining* adalah *topic modelling*. *Topic modelling* adalah satu varian dari *statistical modeling* yang digunakan untuk secara otomatis menemukan topik bahasan dari suatu kalimat, paragraf maupun dokumen [12]. Salah satu algoritma yang banyak digunakan untuk *topic modelling* adalah *Latent Dirichelet Allocation* (LDA). LDA merupakan *unsupervised machine learning technique*, yang bertujuan untuk memberikan ciri dari sebuah dokumen berdasarkan distribusi topik pada suatu korpora [13]. Blei [14] memperkenalkan LDA sebagai model probabilistik generatif untuk kelompok data diskrit seperti teks korpora (kumpulan teks). Dalam eksperimen yang dilakukan dalam penelitian ini, digunakan tiga proses generatif untuk menjalankan LDA yang diadopsi dari penelitian yang dilakukan oleh Clint dan Doss [15], sebagai berikut:

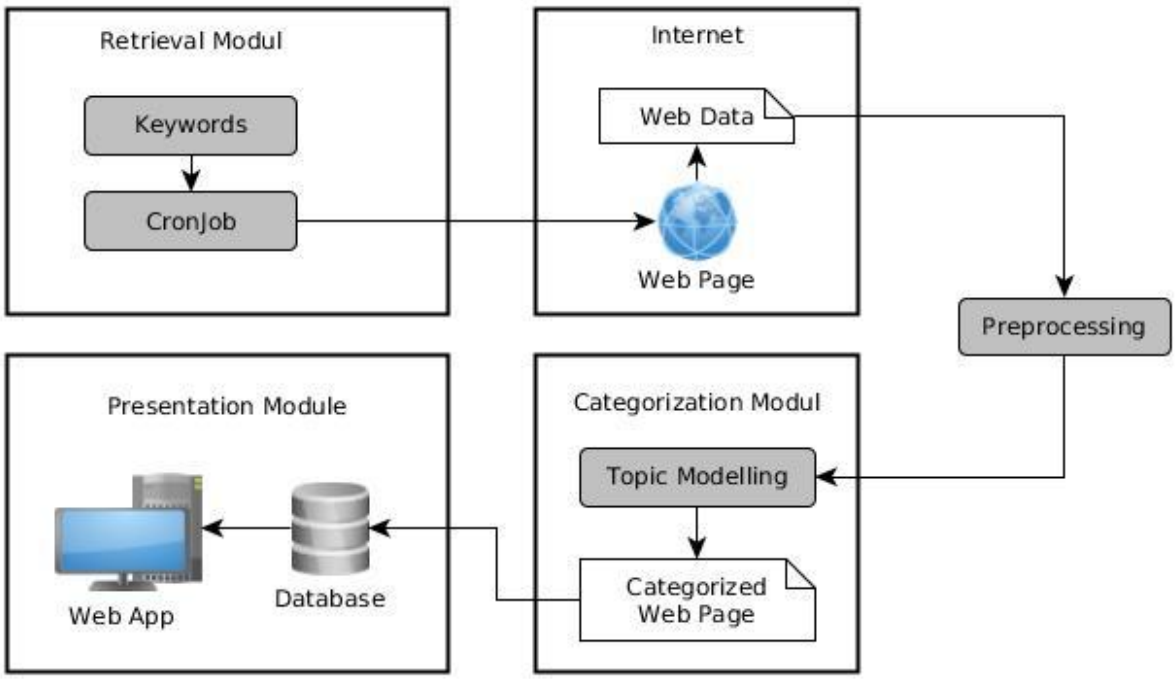
1. Pilih topik secara acak dari distribusinya mengenai topik untuk setiap dokumen.

2. Sampel sebuah kata dari distribusi mengenai kata-kata yang berhubungan dengan suatu topik.
3. Mengulangi proses untuk semua kata dalam dokumen.

4. Hasil dan Pembahasan

Gambar 1, menunjukkan arsitektur sistem yang dikembangkan dalam penelitian ini. Arsitektur yang dikembangkan terdiri dari beberapa modul sebagai berikut:

1. *Retrieval Modul.* Modul ini berfungsi untuk melakukan pengambilan informasi dari laman *web* yang sudah ada. Data dikumpulkan dengan *scraper* yang dapat dibuat dengan berbagai macam bahasa pemrograman. Data dapat dikumpulkan secara berkala dengan memanfaatkan fasilitas *cronjob* yang ada di sistem operasi.
2. *Preprocessing Modul.* Modul ini berfungsi untuk membersihkan data yang didapatkan. P*reprocessing* dilakukan untuk menghindari data yang kurang sempurna, gangguan pada data, dan data-data yang tidak konsisten [16].
3. *Categorization Modul.* Pada modul ini laman *web* yang sudah didapatkan dikelompokkan berdasarkan jumlah kemunculan kata kunci pada laman *web* yang bersangkutan (*term frequency*).
4. *Presentation Modul.* Bagian akhir dari arsitektur ini adalah membangun sebuah laman *web* baru yang menyajikan informasi-informasi yang suah diperoleh dengan *web scraping* dan dikategorisasikan dengan menggunakan *text mining*.

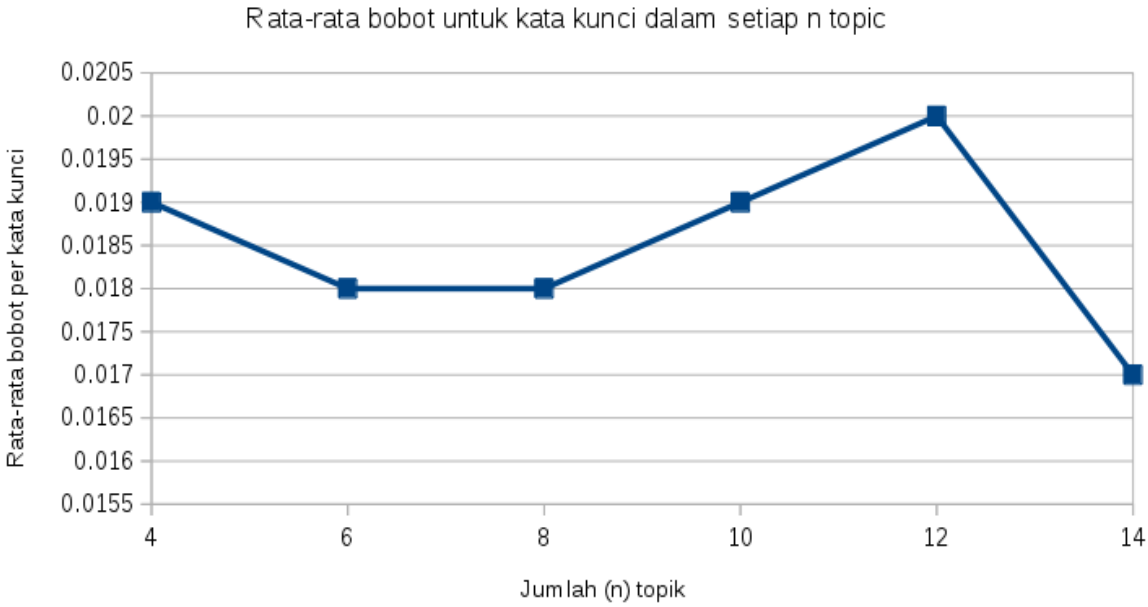


Gambar 1. Arsitektur sistem

Proses awal dari alur arsitektur sistem yang dikembangkan pada Gambar 1 adalah pengambilan informasi dari laman *web* yang sudah ditentukan menggunakan teknik *web scraping*. Dalam menjalankan program *web scraping*, terlebih dahulu harus diketahui struktur HTML dari laman *web* untuk menentukan dalam *tag* HTML yang mana informasi inti direpresentasikan. Tahapan selanjutnya setelah akuisisi informasi dengan *web scraping* adalah pengelompokan informasi dengan menggunakan *text mining*. Metode yang digunakan dalam pengelompokan informasi ini adalah *topic modelling* dengan algoritma LDA. Dalam penggunaan algoritma LDA, pengguna harus menentukan sendiri jumlah topik yang dikehendaki. Hal tersebut tentu saja menyulitkan untuk mendapatkan jumlah

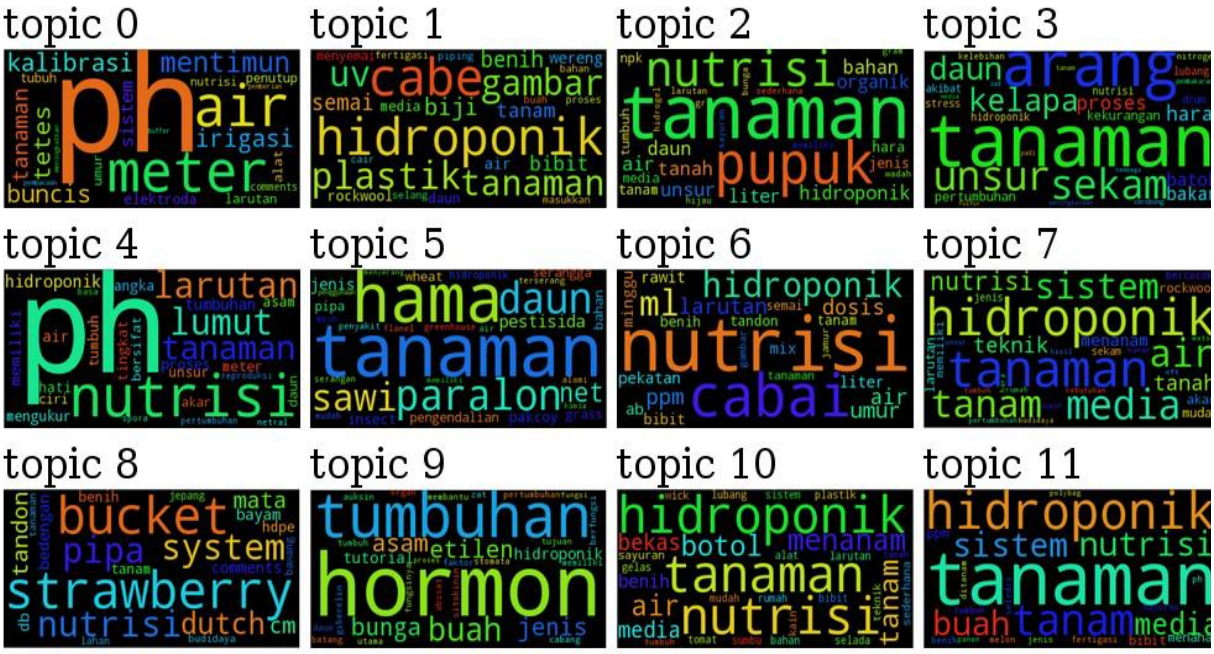
topik yang paling optimal, mengingat pengguna belum tentu seorang *domain expert* yang menguasai topik permasalahan yang diteliti menggunakan LDA.

Untuk mengatasi hal tersebut, di dalam penelitian ini dilakukan eksperimen dengan jumlah topik yang bervariasi mulai dari 4 – 14 topik. Kemudian, kami menghitung nilai rata-rata dari setiap *term*s yang muncul dari setiap model topik. Jumlah topik dengan nilai rata-rata bobot *term* tertinggi kemudian akan dipilih sebagai model topik yang digunakan untuk proses kategorisasi. Bobot rata-rata yang ditunjukkan oleh grafik pada gambar 2 menunjukkan kohesivitas antar topik-topik yang berhasil diidentifikasi dalam korpus artikel hidroponik yang sudah dikumpulkan. Semakin tinggi bobot, maka kohesivitas akan semakin rendah. Kohesivitas antar topik menunjukkan tingkat dependensi atau tingkat *overlapping* antar kata-kata kunci yang menyusun suatu topik pembicaraan. Sehingga topik yang memiliki kohesivitas rendah memiliki komposisi kata-kata kunci yang semakin berbeda/bervariasi dengan topik lainnya. Dari hasil percobaan seperti yang ditunjukkan pada gambar 2, bobot rata-rata tertinggi dimiliki oleh model topik dengan jumlah topik 12. Sehingga, 12 model topik ini yang akan digunakan lebih lanjut untuk melakukan klastering halaman *web* berdasarkan topik yang dikandungnya secara otomatis.



Gambar 2. Grafik rata-rata bobot term yang muncul pada setiap model topik

Langkah selanjutnya adalah memvisualisasikan 12 kelompok topik tersebut dalam bentuk *wordcloud* seperti yang ditunjukkan oleh gambar 3. Visualisasi *wordcloud* dipilih karena dalam visualisasi ini, kata-kata yang memiliki bobot frekuensi kemunculan tinggi dalam suatu topik dapat ditampilkan secara menonjol dari sisi ukuran sehingga lebih intuitif. Dari gambar 3 tersebut terlihat bahwa terdapat beberapa redundansi (memiliki kesamaan/kemiripan kata kunci) yang muncul, seperti topik 0 dengan topik 4, topik 2 dengan topik 10, dlsb. Sehingga untuk efisiensi dan akurasi kategorisasi, maka topik-topik yang redundan tersebut akan dikelompokkan menjadi satu klaster.



Gambar 3. Visualisasi *wordcloud* untuk masing-masing topik

Berdasarkan kemiripan makna dari kata-kata kunci yang muncul di setiap topik, maka topik-topik tersebut dikelompokkan ke dalam lima klaster seperti yang diuraikan pada tabel 2. Selanjutnya, dokumen/laman *web* yang sudah terkumpul dan laman baru dikemudian hari akan dikategorisasikan ke dalam satu atau lebih klaster berdasarkan kata kunci yang menjadi ciri dari setiap klaster. Secara garis besar, klaster 1 berisi konten mengenai pengairan dalam hidroponik, klaster 2 membahas media tanam, klaster 3 berisi pembahasan mengenai nutrisi untuk tanaman hidroponik, klaster 4 membahas mengenai hama yang muncul pada tanaman hidroponik, dan klaster 5 membicarakan tentang jenis-jenis tanaman hidroponik.

Tabel 2. Daftar topik dan kata kunci untuk masing-masing topik.

Klaster	Model Topik	Kata Kunci
Klaster 1	Topik 0, Topik 4	ph, air, meter
Klaster 2	Topik 1, Topik 3, Topik 5, Topik 7, topik 11	plastik, sekam, arang, paralon, media tanam
Klaster 3	Topik 2, Topik 10	nutrisi, pupuk,
Klaster 4	Topik 5	Hama, daun
Klaster 5	Topik 6, Topik 8, Topik 9	Tumbuhan, hormone, cabai, strawberry

5. Kesimpulan

Teknologi *web scraping* dan *text mining* akan memberikan dampak yang sangat bagus apabila diterapkan dalam sistem manajemen pengetahuan berbasis komputer. Kontribusi utama dua teknologi tersebut adalah mengotomasi proses akusisi informasi khususnya informasi-informasi yang bersumber dari artikel atau tulisan bebas di internet. Hal-hal yang akan dilakukan dalam penelitian ini meliputi tiga hal. Hal pertama yang dilakukan adalah melakukan proses pengambilan informasi laman *web* yang ada dengan teknologi *web scraping*. Kemudian, selanjutnya informasi yang berhasil dikumpulkan dikelompokkan kedalam beberapa kategori secara otomatis menggunakan *text mining*. Hasil dari penelitian ini adalah data mengenai artikel hidroponik yang tersebar di berbagai laman *web*

dapat diakuisisi secara otomatis kemudian dikelompokkan menjadi 12 topik menggunakan algoritma *topic modelling* yakni *Latent Dirichelet Allocation* (LDA).

Eksperimen yang dilakukan dalam penelitian ini baru sebatas eksperimen dasar untuk akuisisi dan pengelompokan informasi. Beberapa hal yang dapat dilakukan untuk mengembangkan hasil dari penelitian ini diantaranya adalah:

1. Membuat sebuah portal/laman *web* yang memanfaatkan model topik yang sudah dibuat dalam penelitian ini sehingga bisa secara *real-time* mengumpulkan dan mengkategorisasikan informasi yang muncul setiap saat di internet.
2. Mengeksplorasi algoritma *topic modelling* berdasarkan kesamaan *semantic/ontology* sehingga menghasilkan kategorisasi topik yang lebih akurat. Algoritma LDA yang digunakan dalam penelitian ini masih sebatas melakukan pengelompokkan data berdasarkan *term frequency* dalam suatu korpus dokumen.

6. Ucapan Terima Kasih

Terima kasih dan apresiasi setinggi-tingginya penulis sampaikan kepada pihak-pihak yang membantu terlaksananya penelitian ini. Yang pertama kepada Laboratorium Komputer Fakultas Teknik dan Teknologi Informasi Universitas Jenderal Achmad Yani Yogyakarta yang berkenan meminjamkan komputer *server* untuk proses pengambilan/*crawling* data, serta yang kedua kepada rekan-rekan penulis yang telah memberikan *review* dan masukan atas penelitian yang dilakukan.

7. Referensi

[1] R. C. Wijaya, J. Andjarwirawan, and H. N. Palit, "Aplikasi Pencarian Produk Jual Mobile Devices dari Berbagai Situs E-commerce," *J. Infra*, vol. 4, no. 1, pp. 180–185, 2016.

[2] M. R. Ma'arif, "Integrasi Laman Web tentang Pariwisata Daerah Istimewa Yogyakarta Memanfaatkan Teknologi Web Scraping dan Text Mining," *Teknomatika*, vol. 9, no. 1, pp. 71–80, 2016.

[3] S. Kadam, "Price Comparison of Computer Parts Using Web Scraping," *Int. J. Eng. Sci.*, 2018.

[4] B. G. Dastidar, D. Banerjee, and S. Sengupta, "An Intelligent Survey of Personalized Information Retrieval using Web Scraper," *I.J. Educ. Manag. Eng.*, 2016.

[5] F. Johnson and S. K. Gupta, "Web Content Mining Techniques: A Survey," *Int. J. Comput. Appl.*, vol. 47, no. 11, pp. 44–50, 2012.

[6] M. Turland, *PHP-Architect's Guide to Web Scraping*. Marco Tabini & Associates, 2010.

[7] M. Inzalkar and J. Sharma, "A Survey on Text Mining-techniques and application," *Int. J. Res. Sci. Eng.*, 2015.

[8] S. H. Liao, P. H. Chu, and P. Y. Hsiao, "Data mining techniques and applications - A decade review from 2000 to 2011," *Expert Systems with Applications*. 2012.

[9] S. V. Gaikwad, P. D. Y Patil, and P. Patil, "Text Mining Methods and Techniques," *Int. J. Comput. Appl.*, 2014.

[10] R. S. Pressman, *Software Engineering A Practitioner's Approach 7th Ed - Roger S. Pressman*. 2009.

[11] A. Fathan Hidayatullah, M. Rifqi Ma'arif, "Penerapan Text Mining dalam Klasifikasi Judul Skripsi," *Semin. Nas. Apl. Teknol. Inf. Agustus*, 2016.

[12] S. I. Nikolenko, S. Koltcov, and O. Koltsova, "Topic modelling for qualitative studies," *J. Inf. Sci.*, 2017.

[13] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, 2012.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, 2012.

[15] C. P. George and H. Doss, "Principled Selection of Hyperparameters in the Latent Dirichlet Allocation Model," *J. Mach. Learn. Res.*, 2018.

[16] I. Hemalatha, D. G. P. S. Varma, and D. A.Govardhan, "Preprocessing The Informal Data for Efficient Sentiment Analysis," *Int. J. Emerg. Trends Technol. Comput. Sci.*, 2012.